Stamm, T., Baker, S. R., Foster Page, L. A., Thomson, W. M., Benson, P., Broomhead, T., Aguilar-Diaz, F., Do, L., Gibson, B. J., Hirsch, C., Marshman, Z., McGrath, C., Mohamed, A., Robinson, P. G., Traebert, J., Turton, B., Salzberger, T., & Bekes, K. (2019). Rasch model of the Child Perceptions Questionnaire in multi-country data. *Journal of Dentistry*, [103267]. https://doi.org/10.1016/j.jdent.2019.103267

Peer reviewed version

Link to published version (if available):
10.1016/j.jdent.2019.103267

Rasch model of the Child Perceptions Questionnaire in multi-country data

Revised version R1

Abstract: number of words 250; Clinical significance: number of words 49; body of text (excluding Abstract, Acknowledgments, References, Figures and Tables): number of words 3443; number of tables: 3; Number of figures: 1

Keywords: patient-reported outcome measures, measurement accuracy, cross-border care, psychometric properties, item response theory

Abstract

Objective: To be fit-for-purpose, oral health-related quality of life instruments must possess a range of psychometric properties which had not been fully examined in the 16-item Short Form Child Perceptions Questionnaire for children aged 11 to 14 years (CPQ_{11-14} ISF-16). We used advanced statistical approaches to determine the CPQ's measurement accuracy, precision, invariance and dimensionality and analyzed whether age range could be extended from 8 to 15 years.

Methods: Fit to the Rasch model was examined in 6648 8-to-15-year-olds from Australia, New Zealand, Brunei, Cambodia, Hong Kong, Malaysia, Thailand, Germany, United Kingdom, Brazil and Mexico.

Results: In all but two items, the initial five answer options were reduced to three or four, to increase precision of the children's selection. Items 10 (*Shy/embarrassed)* and 11 (*Concerned what others think)* showed an 'extra' dependency between item scores beyond the relationship related to the underlying latent construct represented by the instrument, and so were deleted. Without these two items, the CPQ was unidimensional. The three oral symptoms items (4 *Food stuck in teeth*, 3 *Bad breath* and 1 *Pain*) were required for a sufficient person-item coverage. In three out of 14 items (21%), Europe and South America showed regional differences in the patterns of how the answer options were selected. No differential item functioning was detected for age.

Conclusion: Except for a few modifications, the present analysis supports the combination of items, the cross-cultural validity of the CPQ with 14 items and the extension of the age range from 8 to 15 years.

Clinical significance**:** The valid, reliable, shortened and age-extended version of the CPQ resulting from this study should be used in routine care and clinical research. Less items and a wider age range increase its usability. Symptoms items are needed to precisely differentiate between children with higher and lower quality of life.

Introduction

Patient-reported outcomes constitute an essential part of health outcome measurement. Valid and precise instruments are a substantial requirement. Health-related quality of life is among the most important domains that can be assessed using self-reported outcome measures; oral health-related quality of life (OHRQoL) is one of its core elements. Self-reported instruments for measuring OHRQoL have been specifically developed for children and adults. The most frequently used instrument in children is the Child Perceptions Questionnaire$_{11\text{-}14}$ (CPQ$_{11\text{-}14}$) [1-3]. The CPQ$_{11\text{-}14}$ was first developed as a 37-item instrument derived from an item pool from different countries and cultures. Items were grouped into four main domains of oral symptoms, functional limitations, emotional and social well-being [2]. Subsequently, a 16-item-short version, the Short Form CPQ$_{11\text{-}14}$ (CPQ$_{11\text{-}14}$ ISF-16), with four items from each of the four main domains was established [4, 5]. A later secondary analysis of data from a subnational sample of 5804 children proposed that the CPQ$_{11\text{-}14}$ worked well with two domains, symptoms/function and emotional/social well-being [5].

To be fit-for-purpose, patient-reported outcome measures must possess a range of psychometric properties that ensure adequate precision and accuracy of measurement, as well as comparability of findings. So far, the psychometric properties of the CPQ have been examined only sufficiently according to classical test theory (CTT) [1, 2, 4, 5]. CTT focuses on overall, sample-based statistics, such as correlations and reliability which provide little insight into how individual items actually work [5]. Furthermore, CTT makes assumptions, such as normally distributed populations and interval-scaled response data, that are rarely met in practice.

An alternative psychometric approach based on the Rasch model for measurement [6] overcomes some limitations of CTT and provides insight at the item level. For this reason, the Rasch model has gained popularity in health measurement in recent years. The Rasch model focuses on the response of an individual to an item which is modeled as a logistic function of item parameters (referred to as the item's difficulty/probability to be affirmed or item location) and a person parameter (the person measure is what we are ultimately interested in). Item and person parameters are expressed in the

same metric scale and are directly comparable. When the fit of the data to the Rasch model and its requirements are investigated, a range of possible misfits might exist. A basic test compares actual responses to the expected responses based on parameter estimates. Other tests check explicitly whether parameter invariance holds true in the data. This is particularly important if measures from potentially different groups of patients (e.g. females versus males or patients from different countries or regions) are to be compared. A violation of invariance is referred to as differential item functioning [7]. It means that the expected response to an item differs between two respondents from different countries even though they have the same person location.

Another fundamental requirement of measurement is unidimensionality. Whenever a response string is summarized by one measure, such as a total (sum) score, unidimensionality is, at least implicitly, assumed meaning that a single underlying latent construct accounts for the variation in the responses of the participants [8]. In concepts comprising multiple aspects or sub-domains, unidimensionality is typically an unrealistic assumption. However, if these aspects are sufficiently related to one another, unidimensionality can still hold true and a single total score adequately measures the latent construct. Related to unidimensionality is local dependency. Local dependency is an 'extra' dependency between item scores beyond the relationship related to the same underlying latent construct, which is measured by the instrument. Local dependency distorts the metric of the measures and is investigated by examining item residual correlations.

Another aspect of a fit-for-purpose instrument is adequate measurement precision; this is referred to as 'targeting' in the Rasch model. It means the extent to which item locations match person locations. In a properly targeted instrument, there is a close match between item and person locations. In poorly targeted instruments, items are too 'easy' (too likely to be observed) or too 'hard' (too unlikely to be observed). A few items that show strong floor or ceiling effects are, generally speaking, unproblematic as they capture extreme person locations. However, if the vast majority of items is affected, the instrument would be poorly targeted to the sample and precision would be very low [9]; the validity of such an instrument is hard to support. Furthermore, the Rasch model allows for a logistic

transformation of an ordinal into a metric scale with a score from 0 to 100. This is particularly relevant in subsequent statistical analyses that require metric data.

Objectives

Except for an initial Rasch analysis using a dataset from Germany [10], the above-described fundamental principles of measurement have not yet been tested for the CPQ_{11-14} ISF-16. Likewise, these principles have not been sufficiently examined in other oral health questionnaires [11, 12]. To date, little is known to what extent the instrument works in different countries and whether measures are cross-nationally comparable. Accordingly, the aim of our study was to use the Rasch model approach to investigate the psychometric properties (fit, invariance, unidimensionality) of the CPQ_{11-14} ISF-16 in a diverse set of countries comprising Australia, New Zealand, Brunei, Cambodia, Hong Kong, Malaysia, Thailand, Germany, United Kingdom, Brazil and Mexico. Another aim was to examine the potential for extending the age range of the CPQ_{11-14} ISF-16 from 8 to 15 years. Such an extension would enhance the clinical usefulness of the instrument and facilitate longitudinal assessments.

Materials and methods

A psychometric analysis was conducted using multi-national epidemiological samples of 6648 children aged 8 to 15 years who completed the CPQ_{11-14}. in 11 countries covering the regions of Australia/New Zealand, Europe, Asia and South America (Supplementary Appendix Figure I). Three hundred seventy-two children were from Australia (data collected in 2002/3; 8-to-13-year-olds), three samples (with 352, 202 and 429 children) from New Zealand (data collected in 2008/10; 12-to-13-year-olds), 423 from Brunei (data collected in 2010; 10-to-14-year-olds), 423 from Cambodia (data collected in 2012; 8-to-14-year-olds), 542 from Hong Kong (data collected in 2001; 12-year-olds), 439 from Malaysia (data collected in 2007; 12-to-13-year-olds), two samples (261, 506) from Thailand (data collected in 2009/11; 10-to-14-year-olds), two samples (88, 374) from UK (data collected in 2003/7/8; 11-to-14-year-olds), 1498 from Germany (data collected in 2007/8; 10-to-15-year-olds), 335 from Mexico (data collected in 2007; 12-to-13-year-olds) and 404 from Brazil (data collected in 2009; 11-to-14-year-olds) (Supplementary Appendix 1). All but the Cambodian and two UK samples were representative at the

national or regional level. Boys (n=3277, 49%) and girls (n=3371, 51%) were represented almost equally. All studies had used either the $CPQ_{11\text{-}14}$ ISF-16 [4] or the 37-item version [2] which also includes the 16 short form items. Response options and scores for each item were as follows: 'Never' (scoring 0); 'Once or twice' (1); 'Sometimes' (2); 'Often' (3); and 'Every day or almost every day' (4).

Fit to the Rasch measurement model

Overall and item-based fit to the Rasch model was explored in a series of analyses using partial credit models suitable for polytomous response data [13]. We used raw scores without weighting [14] to precisely calibrate the scale and to transform the raw scores into a metric scale. As fit statistics are inflated by high sample sizes and simulation studies [15, 16] revealed that sample sizes of 500 appear to be optimal (i.e., not too sensitive while still sufficiently powerful), we analyzed individual item fit in a region-stratified random sample of 125 participants per region (500 in total); we repeated this analysis three times with a different, independent random draw to validate our findings and to check for parameter invariance. This approach has been used in recent, similar studies [17, 18]. Furthermore, we examined for each item whether the locations of the thresholds between the response options were properly ordered. Disordered thresholds indicate that the response scale does not work as intended. In the event of disordered thresholds, we first rescored each item by collapsing the five answer options either into four or three categories (whichever pattern fitted better depending on the inspection of the category probability curves as well as the clinical meanings of the items). Local dependency between items was assessed using residual correlations based on a cut-off of 0.2 above the mean [19].

To assess the instrument's item-based internal consistency and reliability, we compared Cronbach's alpha with the person separation index (PSI). The PSI refers to the reproducibility of relative measure location and indicates whether a scale is able to distinguish between people with higher and lower oral health related quality of life [9]; a PSI ≥0.7 indicates that the instrument is suitable for group comparisons, whereas a PSI value ≥0.85 demonstrates a good person separation for individual use.

Misfitting and locally dependent items were deleted, if their overall fit statistics, reliability measures and clinical meaningfulness based on the information gained from these items were not violated.

Unidimensionality

To test unidimensionality, we used an approach proposed by Smith [20] and combined principal component analysis of the item residuals with a series of t-tests to assess whether subsets of residuals which loaded positively or negatively resulted in different estimates of person parameters. These sets of items were chosen as a way to maximize the contrast between them. These item sets were then most likely to violate the assumption of unidimensionality.

Differential item functioning

Countries were collapsed into the following four regions, namely Australia/New Zealand, Europe, Asia and South America. Differential item functioning was assessed for region and gender, separately for each item by comparing person parameter estimates between different regions. If differential item functioning was apparent for an item, we determined the nature of those differences occurred using post hoc analysis of the residual means. Due to the heterogeneity of the age ranges covered in the datasets from the different countries and the fact that the full age range was not covered in all countries, we assessed differential item functioning for age in an age-stratified random sub-sample of 240 children (30 children for each age group; in years) and repeated this analysis three times using each time a different, independent random draw.

Person-item targeting

Person-item targeting was inspected graphically using person-item map and person-item threshold distribution.

Transformation to a metric interval scale

Based on the above described adaptations, the ordinal total $CPQ_{11\text{-}14}$ raw scores were transformed to a metric scale. If differential item functioning existed, we split the specific item to separate the regions which was different from the others. This approach resulted in region-specific transformation scales. All analyses were performed with RUMM2030 and the eRm package in R (www.r-project.org).

Results

Total CPQ_{11-14} ISF-16 raw scores had a mean of 11.5 (SD 8.6), a median of 10 and a range of 0 to 54. The right-skewed distribution of the total CPQ_{11-14} ISF-16 raw sum scores of the German data indicated that the majority of the German population had a score of zero (no self-reported oral health-related problems and high OHRQoL; Supplementary Appendix Figure II) and affected the interpretation of fit statistics and person-item targeting. Mean Decayed-Missing-Filled Teeth (DMFT) scores of the German data were also numerically lower compared to the other three regions (Supplementary Appendix Figure III). For these reasons, analyses were conducted with and without the data from Germany.

Diagnosis of measurement problems and scale repair

Model fit statistics first showed a considerable initial misfit of the data to the Rasch model, a discrepancy between PSI and Cronbach's alpha, and a significant violation of unidimensionality in the total data set (model 1 in Table 1). Furthermore, in model 1, all but two items had significantly deviating F-tests with item-based fit residuals being below -2.5 or above +2.5 and all but one item showed significantly deviating chi-squared values; all except two items (Item 2 *Sores* and Item 4 *Food stuck in teeth*) produced disordered thresholds. We therefore collapsed and rescored the answer options as depicted in Table 2 in the column named 'rescored'.

Thereafter, model 2 was fitted without the German data (n= 5150). However, this model also showed a considerable misfit (Table 1) with significantly deviating item-based F-tests and chi-squared values. To achieve a better model fit, we drew a smaller region-stratified random sample (model 3 in Table 1) and repeated this procedure three times. We examined the results and similar item locations were produced which are depicted in the Supplementary Appendix Table A. Consequently, item fit residuals in model 3 were numerically smaller (Table 2).

Since local dependency was detected in two items, namely Item 10 *Shy/embarrassed* and Item 11 *Concerned what others think*, it was decided to first delete one of those two items to lose as little information as possible. Because the respective remaining item produced further local dependences with other items, we decided to delete both items. After this procedure, six items still had significantly

deviating F-test and chi-squared values (marked in bold letters in Table 2), despite their item fit residuals being in an acceptable range (between +2.5 and -2.5). Because the deletion of further items also decreased the PSI, we assumed that important discriminative information would be lost and decided not to delete any further items. For this reasons, we also decided not to delete any further items which showed DIF, but to adjust for DIF by generating criterion-specific metric transformation scales. Model 3 showed locally independent items, a unidimensional scale and no differential item functioning by gender. Likewise, none of the age-stratified random samples exhibited differential item functioning for age (Supplementary Appendix Tables B[a] to [c]). However, three items, namely Item 15 *Other kids teased,* Item 8 *Difficulty eating/drinking hot/cold foods* and Item 5 *Taken longer than others to eat*, still had differential item functioning for region in model 3.

When further exploring the differential item functioning for region, the *post hoc* analysis revealed that the children in Europe responded differently to item 8 than in the other three regions. For items 5 and 15, not only Europe, but also South America were different from the other two regions, as well as from each other. Therefore, we split the three items for Europe and South America and transformed the $CPQ_{11\text{-}14}$ ISF-16 raw scores into separate, region-specific metric scales, except for Australia/New Zealand and Asia, which were kept together because no differential item functioning between those two regions was observed in the *post-hoc* analysis (Table 3).

Person item targeting

From the graphical inspection of the person-item map (Figure 1), it is evident that a large number of children had a high probability for a low score, even without the right-skewed German data (lower scores represent better OHRQoL). Furthermore, the three oral symptoms items (Item 4 *Food stuck in teeth*, Item 3 *Bad breath* and Item 1 *Pain*) are needed for a sufficient person-item targeting and to accurately differentiate between children with different levels of OHRQoL. The items on psychosocial consequences showed a similar likelihood to be affirmed to each other and thus, represented only a small proportion of the children. Moreover, including some more new ‘easy’ (more likely to be

affirmed) items on psycho-social consequences would result in an even better person item targeting and a better discrimination by the instrument.

Discussion

Our study used advanced statistical approaches, namely the Rasch model, to investigate the psychometric properties of the CPQ_{11-14} ISF-16 when applied to children from 8-15 years. This is the first study that explored fundamental principles of measurement, including accuracy, precision, invariance and dimensionality in the CPQ_{11-14} ISF-16 covering the regions of Australia/New Zealand, Europe, Asia and South America. Our analysis addressed each item of the instrument and proposed slight improvements that increase its precision and clinical meaningfulness of the questionnaire; with minor adaptations, the CPQ_{11-14} ISF-16 was fit for purpose.

In respect of the dimensionality of the instrument, one recent study investigated the factor structure of the CPQ_{11-14} ISF-16 using confirmatory factor analysis and proposed two subscales (symptoms/function and well-being) [5]. In the same direction, another study in the German dataset suggested the same previously mentioned subscales [10]. Our diagnostic findings also revealed that the unidimensionality of the CPQ_{11-14} ISF-16 was violated. However, after deleting items 10 *Shy/embarrassed* and 11 *Concerned what others think* which showed an 'extra' dependency between the item scores beyond the relationship related to the same underlying latent construct that is measured by the instrument, unidimensionality as a single latent trait was no longer violated. This can be interpreted that the items, factors and/or domains of an instrument follow a common latent trait, namely, in our case, OHRQoL [21]. Consequently, the total score of the instrument has a clinically meaningful interpretation and our findings support, in principle, the successful combination of items of this instrument. For practical purposes, the use of only 14 items instead of 16 would also be more time-efficient for the children. Marshman et al. also found that young people had difficulties to score double items, e.g. item 10 *Shy/embarrassed* [22].

Currently, there are two separate CPQ versions for two age groups, namely the 8–10 and 11–14 ones. However, the use of two measures limits the ability of the CPQ to be used in prospective, longitudinal

studies, that follow individuals throughout childhood. Having a single measure which can be used with children over a ten-year age span would be a considerable advantage [23]. Accordingly, an important finding from our study is that the unified 14-item short form can be used for assessing OHRQoL in a wider age range than has been previously reported. Instead of those two different age-related instruments, our findings support the use of only one unified instrument for both age groups, extending from 8 to 15 years of age. A future approach could be investigating the utility of the 14-item version for adolescents above 15 years of age.

Fayers et al. argued that symptoms were causal indicators and consequently should not be included in OHRQoL instruments [24]. However, our findings suggest that items addressing oral symptoms showed better targeting and coverage of the population than the other items. Thus, oral symptoms items are likely to be an essential part of OHRQoL instruments.

Patient-reported outcome measures should cover what matters to patients, rather than asking only about what health professionals and scientists who developed these instruments considered to be important [25]. Moreover, the frequency, severity and importance of impacts of a health condition should be included in our assessment to capture the value for patients. The Rasch model partly addresses this issue by providing evidence for how the items target the perspectives of the patients who filled in the questionnaire performed. However, further qualitative studies are needed to explore whether the patient-reported outcome measures, including the CPQ, fully cover the perspective of the patients [26-28].

Within all but two items, answer options were collapsed, meaning that the initial five response options were reduced to three or four options, whichever pattern fitted better, to increase the precision of the selection of response options by the children. Collapsing answer options does not necessarily change the layout and format of the questions. It rather represents an algorithm for calculating the total score. The format and layout of the revised version of the CPQ could then look like the current CPQ form - with two items less. Moreover, our findings indicate that the translation and cultural adaptation process was accurate for most items. No differential item functioning was detected in respect to

gender, and only three of the 14 items (21%) showed differential item functioning by region. Those three items were *taken longer than others to eat* (item 5), *difficulty eating/drinking hot/cold foods* (item 8), and *teased by other kids* (item 15). Accordingly, transforming the raw scores into a metric scale specific and separate for each region would allow an accurate comparison of the CPQ_{11-14} scores across international data-sets and when needed, in cross-border care. Moreover, this could facilitate future multi-country studies and support dentists in applying a precise OHRQoL instrument for use with children in their daily practice and/or community setting.

One limitation of this study is that not all samples were representative at the national level. Some data-sets were representative only at regional level or less. Furthermore, the full age range was not covered in all countries. To overcome this limitation, we used region- and an age-stratified random samples for differential item functioning analysis and performed our analysis with and without the data from Germany.

Conclusion

Except for the deletion of two items, the collapsing of the answer options for the calculation of the total score and the region-specific transformation tables, the findings support the combination of items and the cross-cultural validation of the CPQ_{11-14} ISF-16 within the range of included countries. Furthermore, our analysis provides evidence that the CPQ_{11-14} with 14 items is unidimensional and can be used in children aged 8 to 15 years.

Table 1. Model fit statistics. Mean item log residual test of fit, item-trait interaction chi-square statistics, Root Mean Square Error of Approximation (RMSEA) and the Akaike-Information-Criterion (AIC) were calculated to assess model fit.

	Model fit statistics								Unidimensionality analysis			
	Mean item location (± SD)	Mean item fit residual (± SD)	Mean person location (± SD)	Mean person fit residual (± SD)	Person separation index (PSI)	Cronbach's α	Root Mean Square Error of Approximation (RMSEA)	Akaike-Information-Criterion (AIC) - smaller is better	Number of significant t tests	Sample	% PST	Lower bound of 95% CI
Model 1. Total dataset	0	-2.74	-1.58	-0.3	0.79	0.85	0.19	165,166	519	6648	7.8%	7.3%
(± SD)	-0.47	6.35	1.11	1.12								
Model 2. Without the German data	0	-1.17	-1.32	-0.24	0.79	0.83	0.15	156,120	331	5150	6.4%	5.8%
(± SD)	0.45	5.81	1	1.16								
Model 3. 500 region-stratified random sample	0	-0.28	-1.38	-0.24	0.76	0.81	0.04	9,508	23	500	4.6%	2.7%
(± SD)	0.53	1.38	0.97	1.08								

Due to the different sample sizes, differences in AIC between model 1 and 3 as well as between 2 and 3 need to be interpreted with caution.

Table 2. Item fit statistics sorted in a descending order according to item location in the randomly reduced, region-stratified dataset of model 3. If the data fit the Rasch model, the hierarchy of items based on their location parameters can be interpreted. A small (negative) item location implies that the items represent a small amount of the concept of interest ('easy' to affirm items are more likely to be observed and positive responses [affirmations] are in general more likely for 'easy' than for 'hard' items), whereas high (positive) items are 'hard' items in which positive affirmations are more unlikely to be observed. Furthermore, a person with a low person location/estimate is expected to score lower on 'hard' items than a person with a high person estimate. **Fit residuals between -2.5 and +2.5 with non-significant F-tests represented individual item fit. Non-significant chi-squared values were interpreted as fit to the latent trait.** Significances are highlighted in bold letters. * indicates trend/borderline: item 15 had a significant F-test value as well as a chi-squared p-value of 0.073.

Item number		Location	Standard error	Fit residual	F-stat	p-value	ChiSq	p-value	Rescored	DIF
4	*Food stuck in teeth*	-0.83	0.05	1.66	**2.52**	**0.015**	**17.21**	**0.016**		
3	*Bad breath*	-0.81	0.06	2.63	1.42	0.197	11.26	0.128	0-1-2-3-3	
1	*Pain*	-0.50	0.06	-0.69	0.73	0.643	5.15	0.642	0-1-2-3-3	
5	*Taken longer than others to eat*	-0.39	0.08	0.86	0.83	0.561	6.17	0.52	0-1-1-2-2	region
9	*Irritable/frustrated*	-0.26	0.06	-1.72	**3.85**	**>0.001**	**22.7**	**0.002**	0-1-2-3-3	
8	*Difficulty eating/drinking hot/cold foods*	-0.19	0.08	-0.31	1.85	0.077	12	0.1	0-1-1-2-2	region
14	*Argued with other kids*	-0.07	0.08	-0,21	0.55	0.801	3.93	0.788	0-1-1-2-2	
12	*Been upset*	0.14	0.08	-2.07	**4.50**	**>0.001**	**25.69**	**0.001**	0-1-1-2-2	
6	*Difficulty chewing*	0.22	0.09	-1.18	1.8	0.085	11.35	0.124	0-1-1-2-2	
2	*Sores*	0.25	0.06	1.26	**2.22**	**0.031**	**15.97**	**0.025**		
15	*Other kids teased*	0.3	0.09	-1.13	**2.05**	**0.047**	12.95*	0.073	0-1-1-2-2	region
7	*Difficulty saying words*	0.5	0.09	-0.65	0.74	0.642	5.34	0.619	0-1-1-2-2	
13	*Avoided smiling/laughing*	0.7	0.09	-0.88	0.64	0.721	4.75	0.69	0-1-1-2-2	
16	*Other kids asked questions about teeth*	0.94	0.10	-1.49	**3.04**	0.003	**17.37**	0.015	0-1-1-2-2	
10	*Shy/embarrassed*	*Deleted due to local dependency with item 11*								
11	*Concerned what others think*	*Deleted due to local dependency with item 12 and 13 (after deleting item 10)*								

Table 3. Transformation of raw scores into region-specific metric scales. Similar to the findings of our recent study [10], $CPQ_{11\text{-}14}$ total raw scores from Europe might not be precisely transformable to a metric scale at the lower end of the scores due to their right skewed distribution.

Raw score	*Logit* Australia. New Zealand and Asia	*Transformed* Australia. New Zealand and Asia	*Logit* Europe	*Transformed* Europe	*Logit* South America	*Transformed* South America
0	-4.249	0	-4.391	0	-4.235	0
1	-3.441	10	-3.576	9	-3.428	10
2	-2.884	16	-3.01	16	-2.872	16
3	-2.5	21	-2.616	20	-2.489	21
4	-2.199	24	-2.307	24	-2.19	24
5	-1.948	27	-2.047	27	-1.941	27
6	-1.73	30	-1.822	29	-1.725	30
7	-1.535	32	-1.619	32	-1.533	32
8	-1.357	34	-1.434	34	-1.357	34
9	-1.192	36	-1.263	36	-1.195	36
10	-1.038	38	-1.103	37	-1.043	38
11	-0.891	40	-0.951	39	-0.9	39
12	-0.751	41	-0.806	41	-0.762	41
13	-0.615	43	-0.666	42	-0.63	43
14	-0.484	44	-0.53	44	-0.501	44
15	-0.354	46	-0.397	45	-0.375	46
16	-0.227	47	-0.267	47	-0.251	47
17	-0.1	49	-0.137	48	-0.128	48
18	0.027	50	-0.008	50	-0.005	50
19	0.154	52	0.122	51	0.118	51
20	0.283	53	0.254	53	0.243	53
21	0.414	55	0.387	54	0.37	54
22	0.548	56	0.523	56	0.499	56
23	0.686	58	0.663	58	0.633	57
24	0.829	60	0.809	59	0.772	59
25	0.979	62	0.96	61	0.917	61
26	1.137	63	1.12	63	1.071	63
27	1.305	65	1.29	65	1.234	65
28	1.486	67	1.473	67	1.411	67
29	1.684	70	1.674	69	1.606	69
30	1.906	72	1.897	72	1.824	72
31	2.162	75	2.155	75	2.076	75
32	2.468	79	2.463	78	2.38	78
33	2.861	84	2.858	83	2.771	83
34	3.429	90	3.428	89	3.338	89
35	4.253	100	4.253	100	4.163	100

Figure legends

Figure 1. Person item map from the randomly reduced, region-stratified dataset of model 3. The grey bars in the top of the graph refer to the frequencies of the estimated levels of the oral health-related quality of life of the children (person parameters). The black line for each item shows the range of person parameters that this item 'covers'. The numbers below the lines refer to the thresholds between the answer options. The black dot in each line represents the item location. As we rescored all items with had initially disordered thresholds, only ordered thresholds are shown in this graph.

Supplementary figure legends

Supplementary Appendix Figure I. Data were collected in eleven countries (marked in dark blue).

Supplementary Appendix Figure II. Histogram of the total CPQ sum scores with (a) and without the German data (b). With the German data included, the European data depict a right skewed distribution.

Supplementary Appendix Figure III. Decayed-Missing-Filled Teeth (DMFT) scores in the different regions. The figure depicts means with standard errors.

References

[1] A. Jokovic, D. Locker, B. Tompson, G. Guyatt, Questionnaire for measuring oral health-related quality of life in eight-to ten-year-old children, Pediatric dentistry 26(6) (2004) 512-518.
[2] A. Jokovic, D. Locker, M. Stephens, D. Kenny, B. Tompson, G. Guyatt, Validity and reliability of a questionnaire for measuring child oral-health-related quality of life, Journal of dental research 81(7) (2002) 459-463.
[3] F. Gilchrist, H. Rodd, C. Deery, Z. Marshman, Assessment of the quality of measures of child oral health-related quality of life, BMC oral health 14(1) (2014) 40.
[4] A. Jokovic, D. Locker, G. Guyatt, Short forms of the Child Perceptions Questionnaire for 11–14-year-old children (CPQ 11–14): development and initial evaluation, Health and quality of life outcomes 4(1) (2006) 4.
[5] W.M. Thomson, L.A. Foster Page, P.G. Robinson, L.G. Do, J. Traebert, A.R. Mohamed, B.J. Turton, C. McGrath, K. Bekes, C. Hirsch, F. Del Carmen Aguilar-Diaz, Z. Marshman, P.E. Benson, S.R. Baker, Psychometric assessment of the short-form Child Perceptions Questionnaire: an international collaborative study, Community Dent Oral Epidemiol 44(6) (2016) 549-556.
[6] G. Rasch, On General Laws and Meaning of Measurement in Psychology, Proceedings of the Fourth Berkley Symposium on Mathemtical Statistics and Probability 4 (1961) 321-333.
[7] G. Tutz, G. Schauberger, A penalty approach to differential item functioning in Rasch models, Psychometrika 80(1) (2015) 21-43.
[8] M. Heene, A. Kyngdon, P. Sckopke, Detecting Violations of Unidimensionality by Order-Restricted Inference Methods, Frontiers in Applied Mathematics and Statistics 2 (2016) 3.
[9] A. Tennant, P.G. Conaghan, The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper?, Arthritis care & research 57(8) (2007) 1358-1362.
[10] M. Omara, T. Stamm, M. Boecker, V. Ritschl, E. Mosor, T. Salzberger, C. Hirsch, K. Bekes, Rasch model of the Child Perceptions Questionnaire for oral health–related quality of life: A step forward toward accurate outcome measures, The Journal of the American Dental Association (2019).
[11] P.E. Benson, S.J. Cunningham, N. Shah, F. Gilchrist, S.R. Baker, S.J. Hodges, Z. Marshman, Development of the Malocclusion Impact Questionnaire (MIQ) to measure the oral health-related quality of life of young people with malocclusion: part 2 - cross-sectional validation, J Orthod 43(1) (2016) 14-23.
[12] H.M. Wong, C.P. McGrath, N.M. King, Rasch validation of the early childhood oral health impact scale, Community dentistry and oral epidemiology 39(5) (2011) 449-457.
[13] A. Tennant, J. Pallant, Unidimensionality matters!(A Tale of Two Smiths?). Rasch Measurement Transactions, 20 (1), 1048–1051, 2006.
[14] D.J. Caplan, G.D. Slade, S.A. Gansky, Complex sampling: implications for data analysis, Journal of public health dentistry 59(1) (1999) 52-59.
[15] A.B. Smith, R. Rush, L.J. Fallowfield, G. Velikova, M. Sharpe, Rasch fit statistics and sample size considerations for polytomous data, BMC medical research methodology 8(1) (2008) 33.
[16] P. Hagell, A. Westergren, Sample Size and Statistical Conclusions from Tests of Fit to the Rasch Model According to the Rasch Unidimensional Measurement Model (Rumm) Program in Health Outcome Measurement, Journal of applied measurement 17(4) (2016) 416-431.
[17] M. Armstrong, C. Morris, M. Tarrant, C. Abraham, M.C. Horton, Rasch analysis of the Chedoke-McMaster Attitudes towards Children with Handicaps scale, Disabil Rehabil 39(3) (2017) 281-290.
[18] W.J. Taylor, K. Parekh, Rasch analysis suggests that health assessment questionnaire II is a generic measure of physical functioning for rheumatic diseases: a cross-sectional study, Health and quality of life outcomes 16(1) (2018) 108.
[19] D. Andrich, Rasch models for measurement, Sage1988.

[20] E.V. Smith Jr, Understanding Rasch measurement: Detecting and evaluating the impact of multidimenstionality using item fit statistics and principal component analysis of residuals, Journal of applied measurement (2002).
[21] M.J. Coppes, S.A. Fisher-Owens, Oral Health in Children, An Issue of Pediatric Clinics of North America E-Book, Elsevier Health Sciences2018.
[22] Z. Marshman, B.J. Gibson, P.E. Benson, Is the short-form Child Perceptions Questionnaire meaningful and relevant to children with malocclusion in the UK?, Journal of orthodontics 37(1) (2010) 29-36.
[23] M. Wilson-Genderson, H.L. Broder, C. Phillips, Concordance between caregiver and child reports of children's oral health-related quality of life, Community dentistry and oral epidemiology 35 (2007) 32-40.
[24] P.M. Fayers, D.J. Hand, K. Bjordal, M. Groenvold, Causal indicators in quality of life research, Quality of life research 6(5) (1997) 393-406.
[25] D. Locker, F. Allen, What do measures of 'oral health-related quality of life'measure?, Community dentistry and oral epidemiology 35(6) (2007) 401-411.
[26] T. Stamm, G.F. Van der, C. Thorstensson, E. Steen, F. Birrell, B. Bauernfeind, N. Marshall, B. Prodinger, K. Machold, J. Smolen, M. Kloppenburg, Patient perspective of hand osteoarthritis in relation to concepts covered by instruments measuring functioning: a qualitative European multicentre study, Ann.Rheum.Dis. 68(9) (2009) 1453-1460.
[27] T.A. Stamm, V. Nell, M. Mathis, M. Coenen, D. Aletaha, A. Cieza, G. Stucki, W. Taylor, J.S. Smolen, K.P. Machold, Concepts Important to People with Psoriatic Arthritis are not Adequately Covered by Standard Measures of Functioning, Arthritis Rheum 57(3) (2007) 487-494.
[28] M.A. Stoffer, J.S. Smolen, A. Woolf, A. Ambrozic, F. Berghea, A. Boonen, A. Bosworth, L. Carmona, M. Dougados, M. de Wit, J. Erwin, V. Fialka-Moser, R. Ionescu, A.M. Keenan, E. Loza, R.H. Moe, R. Greiff, P. Olejnik, I.F. Petersson, A.C. Rat, B. Rozman, B. Strombeck, L. Tanner, T. Uhlig, T.P. Vlieland, T.A. Stamm, W.P.E.P. eumusc.net, Development of patient-centred standards of care for osteoarthritis in Europe: the eumusc.net-project, Ann Rheum Dis 74(6) (2015) 1145-9.